MP 5.4 A 7th-Generation x86 Microprocessor

Steven Hesley, Victor Andrade, Bob Burd, Greg Constant, Jeffrey Correll, Matthew Crowley, Michael Golden, Nancy Hopkins, Saiful Islam, Scott Johnson, Rabbani Khondker, Dirk Meyer, Jerry Moench, Hamid Partovi, Randy Posey, Fred Weber, John Yong

Advanced Micro Devices, Austin, TX

The AMD-K7 (TM) processor is an out-of-order, three-way superscalar x86 microprocessor with a 15-stage pipeline, organized to allow 500+MHz operation. The processor can fetch, decode, and retire up to three x86 instructions per cycle to independent integer and floating-point schedulers. The schedulers can simultaneously issue up to nine operations to seven integer and three floating point execution resources. The cache subsystem and memory interface minimize effective memory latency and provide high bandwidth data transfers to and from these execution resources. The processor contains separate instruction and data caches, each 64kB and two-way set-associative. The data cache is banked and supports concurrent access by two loads or stores, each up to 64b in length. The processor contains logic to directly control an external L2 cache. The L2 data interface is 64b wide and supports bit rates up to 2/3 the processor clock rate. The system interface consists of a separate 64b data bus.

The processor uses a variation of the pulsed flip-flop as the basic latching element. In addition to its small latency (Tsu + Tcq), this topology incorporates complex logic in its first dynamic stage through the nMOS pull down network (PDN), as shown in Figure 5.4.1. To minimize hold time without significantly affecting yield, a statistical model based on local variation of devices is used to determine the smallest CLKPULSE width. The model limits overall yield fallout to 0.1% due to failure to capture data. Master-slave latches in non-critical paths eliminate hold time concerns and reduce power.

The read-only memory (ROM) arrays are self-timed edge-triggered full-rail segmented structures. Each ROM array consists of one, two, or four 64b tall arrays connected by a super bit line. The full rail segmented architecture has comparable speed, power, and smaller area to a reference cell and static load design. The segmented and twisted bit lines reduce the coupling to aggressors. Full rail circuits eliminate the risk and complexity of matching circuits and races associated with small signals.

The sub-1ns static random access memory (SRAM) arrays are single-cycle or pipelined. The single-cycle arrays dynamically decode the address and access the array in the same cycle. The first stage of decode is incorporated in the edge-triggered, self-resetting address flops. These address flops generate monotonic outputs that drive two NAND decoders.

The pipelined arrays are required in the data cache (DC) to support three cycle load latency. Cycle 1 is used by the load store unit. The majority of cycle 2 is for address steering and transport, as shown in Figure 5.4.2. Because there is no clock edge available, and any self-timed signal to match the worst case address path increases the cycle time, the address is statically decoded late in cycle 2. To compensate for the increased area for latching the decoded address, the scan logic and the clock pulse generator are removed from the address flops. A test clock is added for scan controllability of these flops, and the decoder is used as the pulse generator as shown in Figure 5.4.3. The penalty for the pipelined arrays is less than 1% in area and speed.

The processor has two custom register files: the 88-entry, 90b, five read, five write, floating point register file (FPRF) and the 24-entry, 32b, nine read, eight write, combined integer future file and register file (IFFRF). Both register files avoid complex bypass circuitry by completing writes before reads occur. The FPRF decodes the write address in the previous cycle, so that the write naturally completes during read address decode. The IFFRF delays the read access until the write and tag comparisons complete. To reduce the routing and area cost, the write bitlines are single-ended. The low voltage writeability issue is solved by the three transistor configuration used for each write port, as shown in Figure 5.4.4.

The phase-locked loop (PLL) operates with a 2.5V supply, internally regulated down to 1.6V to satisfy oxide voltage stress limits. A high precision bandgap circuit minimizes variation of this internal supply voltage. Given the limited voltage headroom and the high frequency target, the PLL is designed to maximize the voltage controlled oscillator (VCO) control range. To ensure minimum static phase error over the maximum VCO control voltage range, the charge pump is designed to regulate the UP current level based on the DOWN level. This avoids large current mismatches when the UP current source devices begins to exit saturation. The cycle compression (less than 25ps) is optimized at the expense of accumulated phase error (less than 1ns) by setting the loop natural frequency low.

The PLL clock is transported to the center of the chip. From the center of the chip, an eight-level binary tree distributes the clock to eight horizontal buffer slices. The final programmable drivers are connected to the metal-5 and metal-6 mesh grid in 66 columns across each buffer slice. The maximum RC simulated skew, shown in Figure 5.4.5, is 32ps. The simulated process skew, due to channel length variation, is 96ps.

The chip is designed with full scan. One scan chain is dedicated for programming of self-timed pulses in the macros. A 13N march C algorithm is used for the DC arrays, IC arrays, and register files. At low frequency, the DC and IC bit cells are tested for data retention. For debug, the chip also supports on-the-fly frequency variation, single-cycle step operation, and stop mode. For PLL characterization, a scan chain, two high-speed pads, and one analog pad are used for extensive measurements of critical clock phase relationships and sub-block operations.

The die is 1.84cm^2 and contains 22M transistors. Table 5.4.1 shows the technology features. C4 solder-bump flip-chip technology is used to assemble the die into a ceramic 575-pin BGA. Measurements are from initial silicon evaluation, unless otherwise stated.

 0-7803-5129-0/99

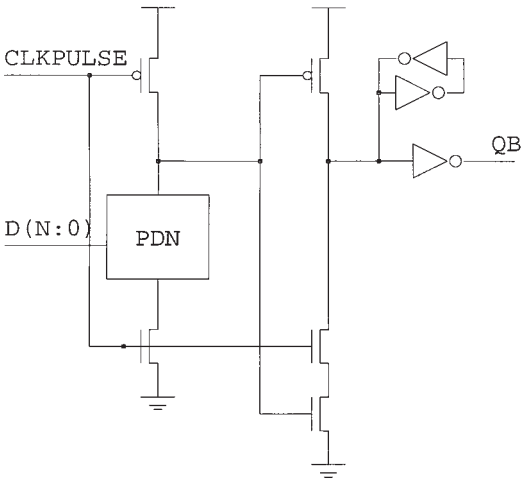

Figure 5.4.1: Basic flip-flop.

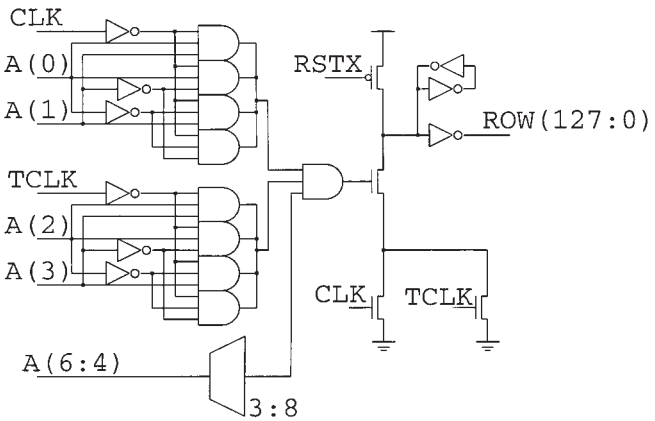

Figure 5.4.3: DC row decoder.

Figure 5.4.5: Clock RC skew.

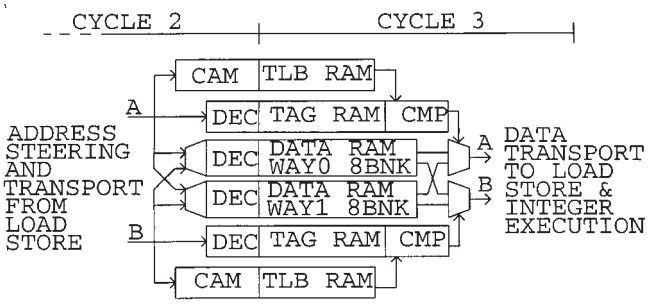

Figure 5.4.2: DC pipeline diagram.

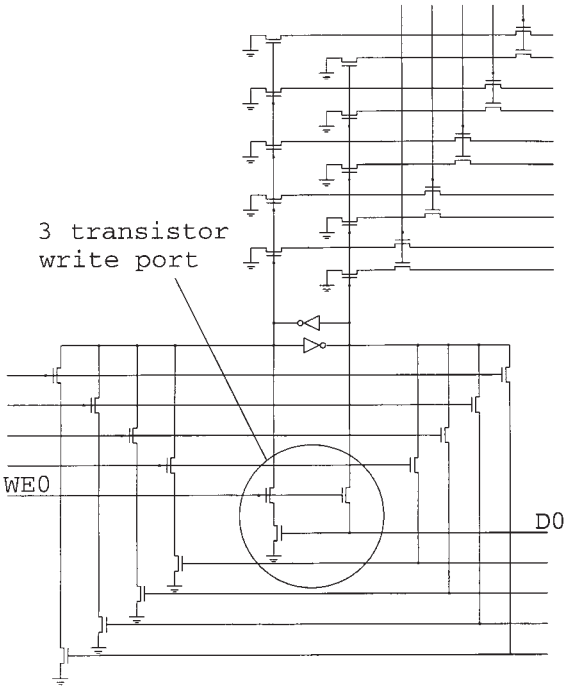

Figure 5.4.4: FPRF bit cell.

Leff	0.16um
Gate Oxide Thickness	3.5nm
Local Interconnect Pitch	0.625um
Metal 1,2 Pitch	0.875um
Metal 3	1.000um
Metal 4,5	1.250um
Metal 6	3.000um
Power Supply	1.60V

Table 5.4.1: Technology features.

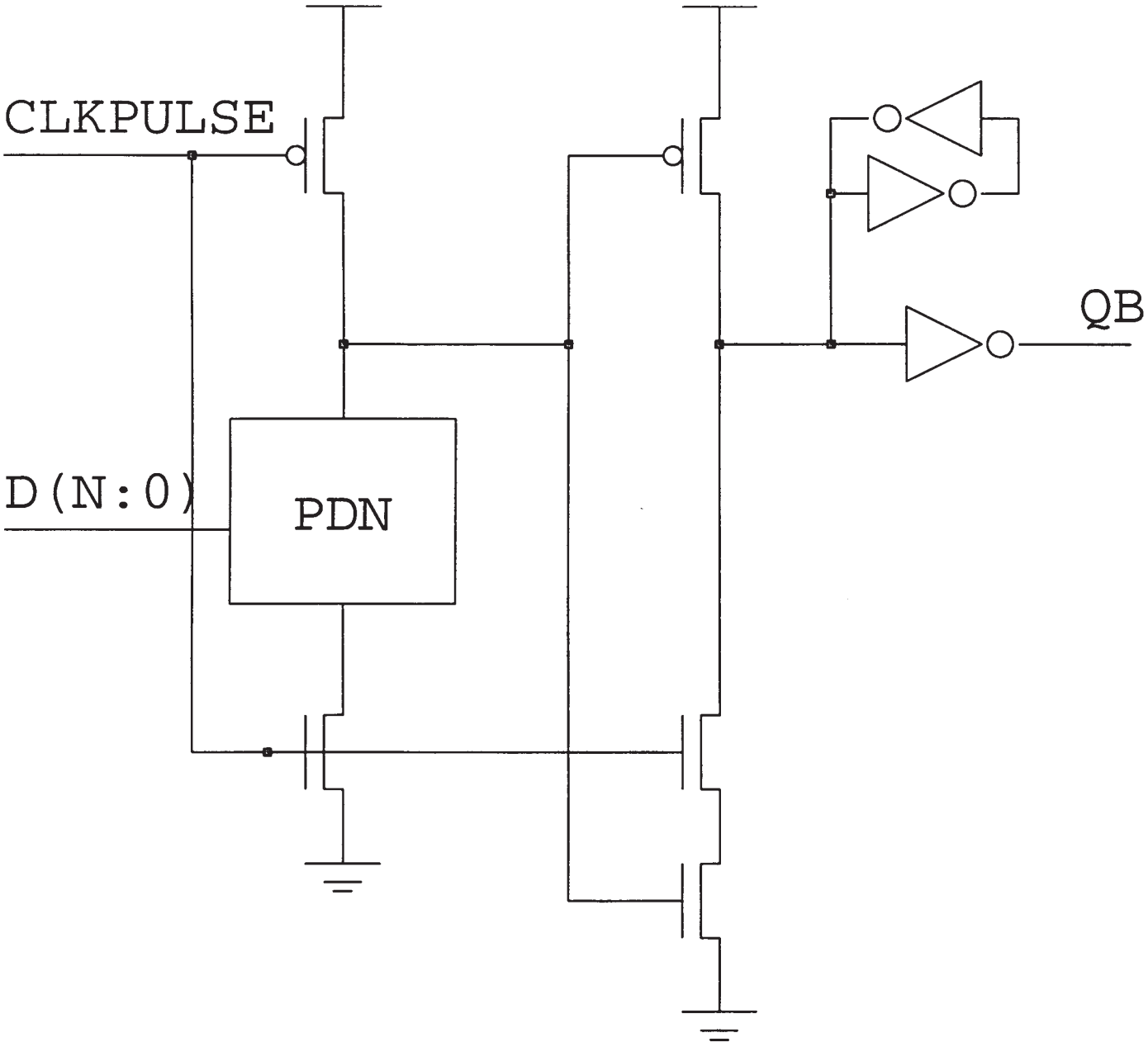

Figure 5.4.1: Basic flip-flop.

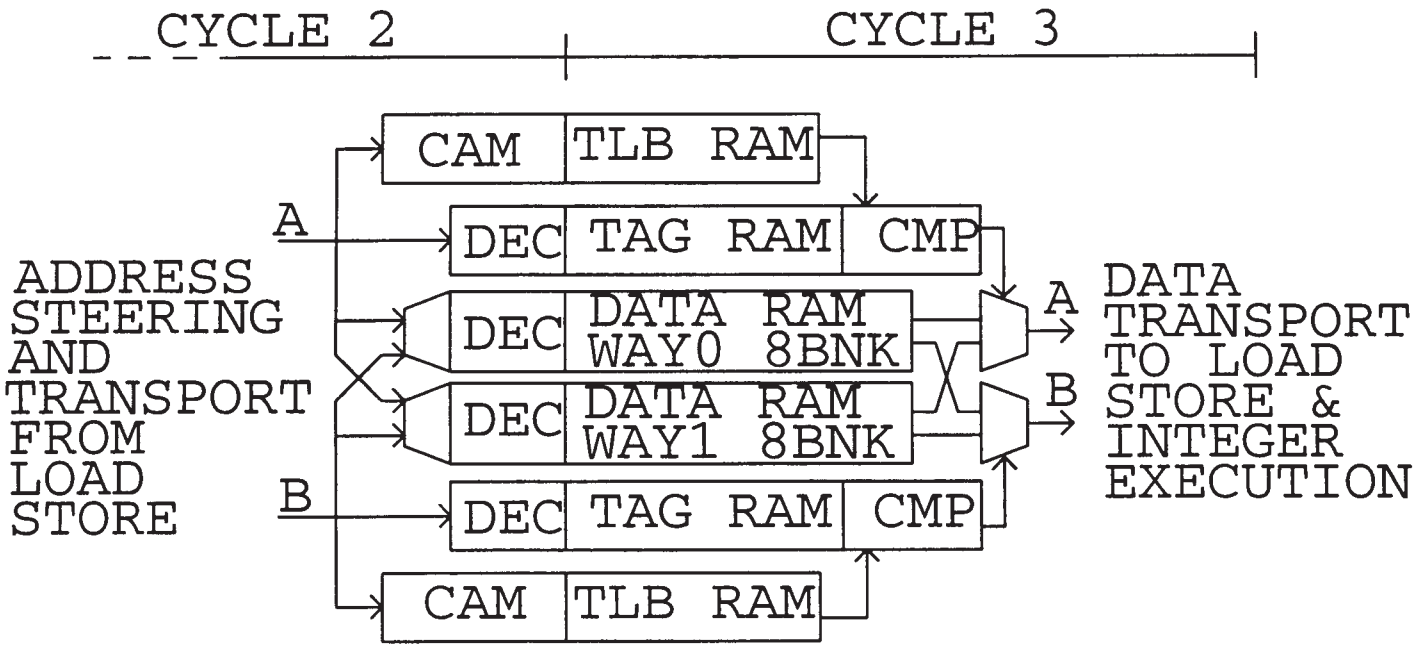

Figure 5.4.2: DC pipeline diagram.

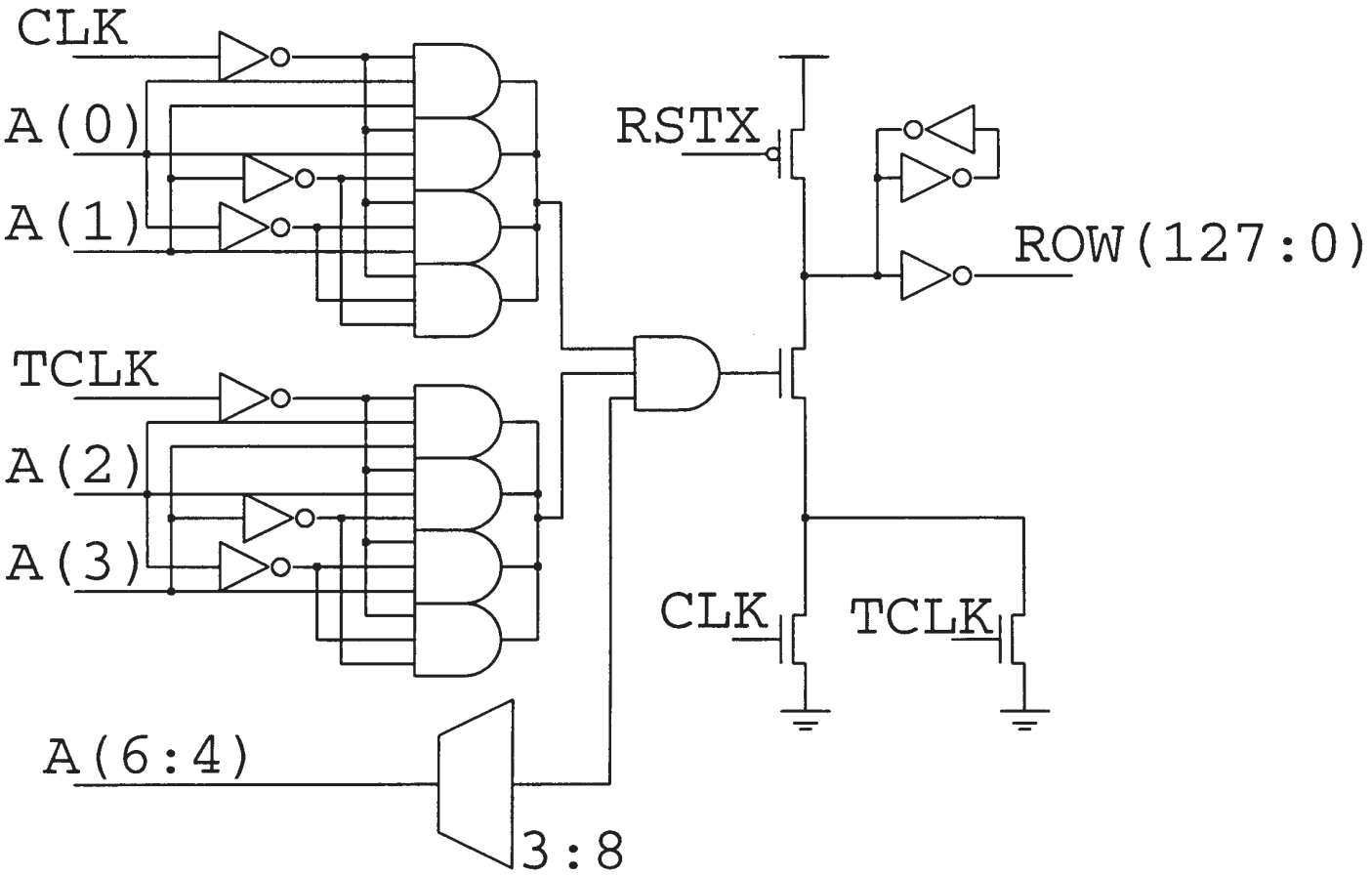

Figure 5.4.3: DC row decoder.

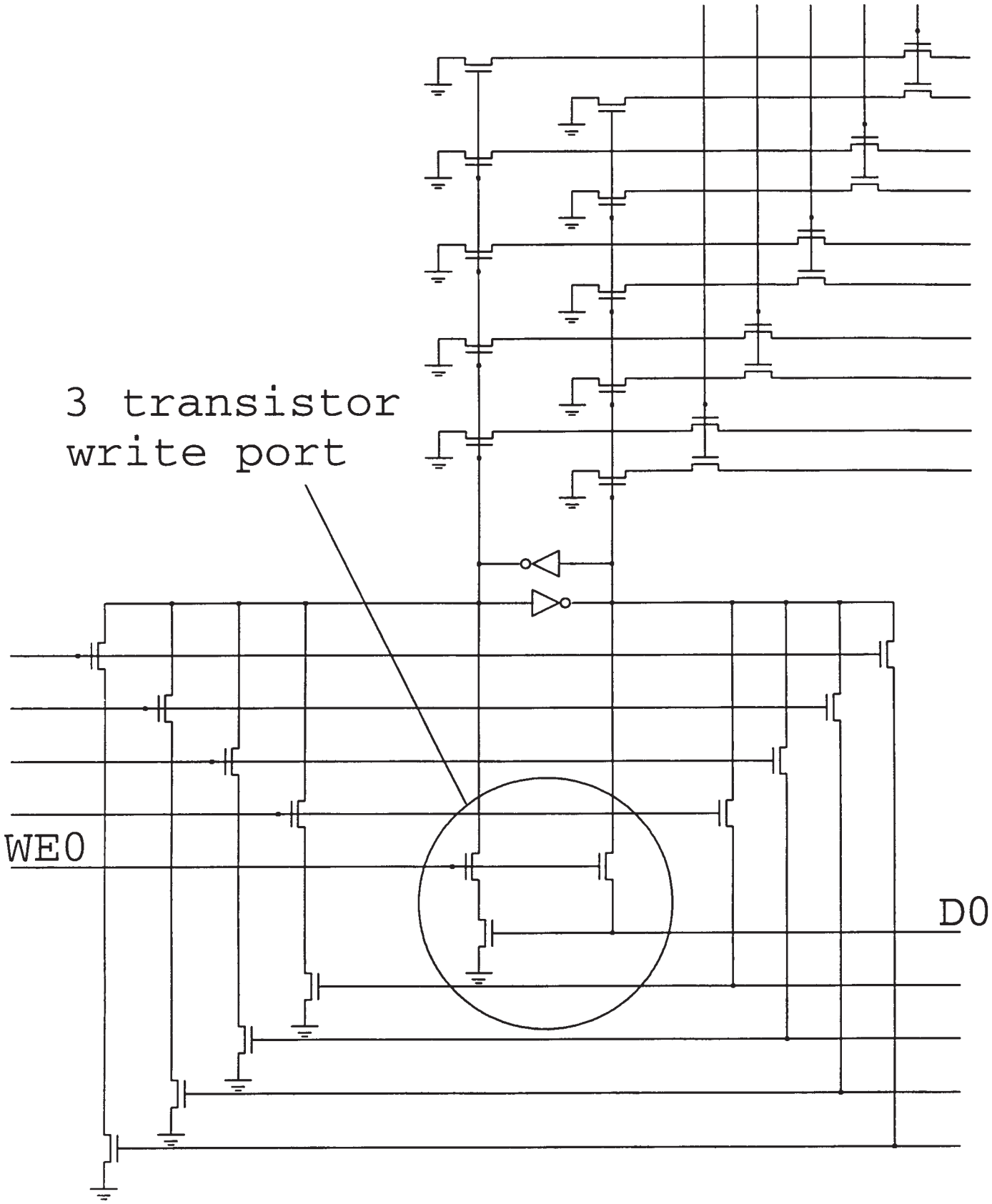

Figure 5.4.4: FPRF bit cell.

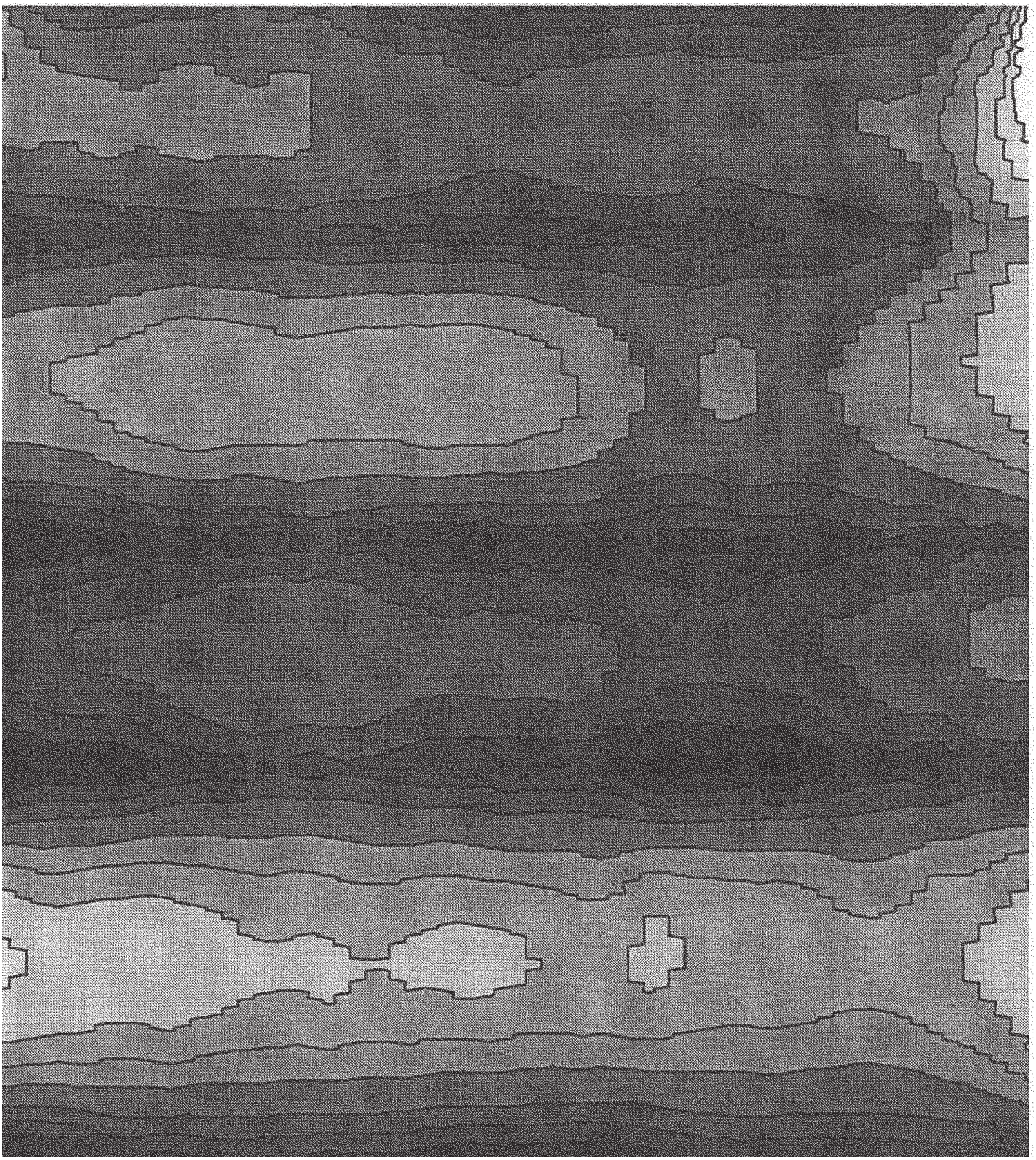

Figure 5.4.5: Clock RC skew.

Leff	0.16um
Gate Oxide Thickness	3.5nm
Local Interconnect Pitch	0.625um
Metal 1,2 Pitch	0.875um
Metal 3	1.000um
Metal 4,5	1.250um
Metal 6	3.000um
Power Supply	1.60V

Table 5.4.1: Technology features.

